JOURNAL OF BACTERIOLOGY, Aug. 1986, p. 639–646
0021-9193/86/080639-08$02.00/0
Copyright © 1986, American Society for Microbiology

Vol. 167, No. 2

crp Genes of *Shigella flexneri*, *Salmonella typhimurium*, and *Escherichia coli*

PASCALE COSSART,[1]* EDUARDO A. GROISMAN,[2] MARIE-CLAUDE SERRE,[1] MALCOLM J. CASADABAN,[2] AND BRIGITTE GICQUEL-SANZEY[1]

Unité de Biochimie Cellulaire, Département de Biochimie et Génétique Moléculaire, Institut Pasteur, 75724 Paris Cedex 15, France,[1] and Department of Molecular Genetics and Cell Biology, The University of Chicago, Chigaco, Illinois 60637[2]

Received 12 November 1985/Accepted 8 May 1986

The complete nucleotide sequences of the *Salmonella typhimurium* LT2 and *Shigella flexneri* 2B *crp* genes were determined and compared with those of the *Escherichia coli* K-12 *crp* gene. The *Shigella flexneri* gene was almost like the *E. coli crp* gene, with only four silent base pair changes. The *S. typhimurium* and *E. coli crp* genes presented a higher degree of divergence in their nucleotide sequence with 77 changes, but the corresponding amino acid sequences presented only one amino acid difference. The nucleotide sequences of the *crp* genes diverged to the same extent as in the other genes, *trp*, *ompA*, *metJ*, and *araC*, which are structural or regulatory genes. An analysis of the amino acid divergence, however, revealed that the catabolite gene activator protein, the *crp* gene product, is the most conserved protein observed so far. Comparison of codon usage in *S. typhimurium* and *E. coli* for all genes sequenced in both organisms showed that their patterns were similar. Comparison of the regulatory regions of the *S. typhimurium* and *E. coli crp* genes showed that the most conserved sequences were those known to be essential for the expression of *E. coli crp*.

The catabolite gene activator protein (CAP) is a positive regulator of gene expression. When complexed with cyclic AMP (cAMP), it binds specifically in the promoter region of the operons which it regulates (14, 49). CAP represents a model system for the study of a pleiotropic regulation of gene expression. The sequence of the *Escherichia coli* K-12 *crp* gene coding for CAP (2, 9), and the structure of the CAP-cAMP complex have been determined, the latter from a 0.29-nm-resolution electron density map (29). By using a genetic approach, we have proposed a model for the CAP-DNA-specific interaction (15, 16). Another genetic study has proposed a model describing the effect of cAMP on the activation process by CAP (19). In vivo and in vitro studies have shown that *crp* is autoregulated (1, 10). The transcription start site has been identified, and different regulatory elements, such as the binding sites for CAP and for the RNA polymerase, have been located (1). However, the significance of other features of the nucleotide sequence preceding the structural gene, such as inverted repeats (2) or a 213-base-pair open reading frame (our published data), remains unknown. Our interest in elucidating the role of the structure in CAP function prompted us to analyze the structure of CAP in different bacteria. In addition, by comparing the nucleotide sequences of the regulatory elements of different *crp* genes, we hoped to gain insight into those aspects of the sequence which are functionally important for *crp* expression. By using an in vivo cloning system (21), we cloned the *crp* genes of *Salmonella typhimurium* LT2 and *Shigella flexneri* 2B. In this paper, we report the sequences of these genes and their surrounding regions and present a comparative study of the *E. coli*, *S. typhimurium*, and *Shigella flexneri crp* genes.

MATERIALS AND METHODS

Materials. The media used were described by Miller (32). Nucleotide sequencing reagents ([γ-^{32}P]ATP [specific activity, 3,000 Ci/mmol], [γ-^{32}P]deoxynucleoside triphosphates [3,000 Ci/mmol], and [α-^{35}S]deoxynucleoside triphosphates [specific activity, >400 Ci/mmol]) and nick translation and sequencing kits were obtained from Amersham Corp. Klenow fragments and T4 DNA ligase were acquired from New England BioLabs, Inc.; T4 kinase was obtained from Boehringer Mannheim Biochemicals; restriction endonucleases were acquired from Boehringer or New England BioLabs; and acrylamide was obtained from British Drugs Houses. All enzymes were used according to the instructions of the supplier. Eosin-methylene blue or MacConkey plates were supplemented with ampicillin (50 μg/ml), chloramphenicol (25 μg/ml), or tetracycline (12.5 μg/ml).

Bacterial strains. For the cloning, the starting stains were *S. typhimurium* LT2 SL4213 (13) and *Shigella flexneri* 2B ATCC 12022. Strains carrying *crp*-20B (strain LU53) (41), *crp-45* (strain BS680) (9), and Δ(*crp-45 cya-06*) (strain CA8445) (42) mutations were used as hosts for the plasmids.

Plasmid clones. Plasmid pBC4042, containing the mini-Mu replicon Mu dII4042 and two antibiotic (chloramphenicol and ampicillin) resistance genes, was introduced into *Shigella flexneri* and *S. typhimurium* Mu cts lysogens by transformation and selection for ampicillin and chloramphenicol resistance (21; E. A. Groisman and M. J. Casadaban, manuscript in preparation). Transformants were heated to induce transposition to different sites during phage replication. DNA sequences could then be flanked by copies of Mu dII4042. Packaging started from the left side of one Mu dII4042 and could include bacterial sequences together with the Mu sequences inserted on the other side. After infection with such a lysate, homologous reombination between Mu sequences could take place to form a plasmid carrying the gene to be cloned. For us, the lysates were used to infect an *E. coli crp* mutant strain (42). Plasmid DNA was isolated from the cyclic AMP receptor protein (CRP$^+$) transductants identified as Mal$^+$ (red colonies on MacConkey maltose chloramphenicol plates) and was used to transform another *crp* strain (BS680). We verified that all chloramphenicol-resistant transformants were CRP$^+$ and that all CRP$^+$ trans-

* Corresponding author.

formants were chloramphenicol resistant. Therefore plasmid pEG5077 was presumed to contain the *crp* region from *Shigella flexneri*, and plasmid pEG5032 was thought to contain the *crp* region from *S. typhimurium*.

pBR322 derivative plasmids pSF281 and pSF280 were constructed, respectively, by ligation of pBR322 and pEG5077 cleaved by *Bam*HI and ligation of pBR322 and pEG5032 cleaved by *Pst*I and by transformation of strain BS680. Transformed bacteria were plated on eosin-methylene blue maltose agar supplemented with ampicillin or tetracycline. The CRP$^+$ transformants were identified as maltose$^+$. Strains harboring plasmids containing the subcloned *crp* region from *Shigella flexneri* were selected as Ap^r (pSF281), whereas strains harboring plasmids containing the subcloned *crp* region from *S. typhimurium* were selected as Tc^r (pST280) (see Results).

Plasmid pST278 was constructed as follows. DNA fragments generated by a partial *Sau*3AI digestion of pST280 were ligated with pBR322 DNA cleaved by *Bam*HI. The ligation mixture was used to transform strain BS680. CRP$^+$ transformants were isolated as Ap^r maltose$^+$ colonies, as described above.

The construction of plasmids carrying deletions between identical restriction sites was performed as previously described (9).

DNA hybridization. Southern blots (47) were performed essentially as described by Maniatis et al. (27). Restriction digests of plasmid DNA were separated on agarose gels, denatured, transferred to nitrocellulose filters, and immobilized. The DNA fragments were then hybridized to a ^{32}P-labeled (nick-translated) (39) *Hin*dIII-*Eco*RV DNA fragment containing the *E. coli crp* gene. Autoradiography was used to locate the band complementary to the radioactive probe.

Sequencing. Plasmid DNA was isolated by the method of Birnboim and Doly (5). The DNA fragments were purified from thin polyacrylamide gels and eluted by diffusion in a crush-and-soak buffer (28) or electroeluted. Fragments were 5′ or 3′ end labeled as previously described (9) and sequenced by the procedure of Maxam and Gilbert (28). Sequence determinations were also performed by the method of Sanger et al. (44) with [α-^{35}S]dATP and gradient gels (4) after cloning the *Hin*dIII-*Eco*RV fragment carrying the *Shigella flexneri crp* gene in M13mp9 (30). We used either the 17-mer universal primer or an internal primer kindly given by M. Gent and S. Minter (University of Manchester Institute of Science and Technology, Manchester, United Kingdom). We also randomly cloned in M13mp8 the *Hae*III digest of the *Sal*I-*Bst*EII fragment containing the 5′ end of the *crp* gene of *S. typhimurium*. In that case, we used either the universal primer or a 17-mer internal primer complementary to the translation initiation codon region.

RESULTS AND DISCUSSION

Cloning and sequencing. We cloned the *crp* genes from *S. typhimurium* and *Shigella flexneri* by their ability to complement an *E. coli crp* mutant (41) and by use of an in vivo cloning system already described (see Materials and Methods) (21). We verified that the resulting plasmids harbored a gene analogous to *crp* by DNA-DNA hybridization experiments. Restriction digests of the plasmids were probed with the 740-base-pair *Hin*dIII-*Eco*RV DNA fragment containing 110 base pairs in addition to the structural *crp* gene of *E. coli*. For the plasmid containing an insert from *Shigella flexneri*, we found a positive hybridization with a *Bam*HI fragment of the same size as the *Bam*HI fragment of *E. coli* containing *crp* (9). This fragment was cloned into the *Bam*HI site of pBR322 to give rise to plasmid pSF281. For the plasmid containing an insert from *S. typhimurium*, we found a positive hybridization with a 7,200-base-pair *Pst*I fragment that we subsequently cloned in the *Pst*I site of pBR322 to give rise to plasmid pST280.

Both plasmids complemented the carbohydrate-negative phenotype of a *crp cya E. coli* strain (CA8445) only in the presence of cAMP. These results confirmed those obtained from the hybridization experiment and showed that the recombinant plasmids contained the *crp* genes of *Shigella flexneri* and *S. typhimurium*. The restriction maps of the *crp* regions of *E. coli* and *S. flexneri* were very similar: *Bam*HI, *Hin*dIII, *Bcl*I, *Hpa*I, *Kpn*I, *Sal*I, and *Eco*RV sites were at the same locations in both organisms. There was one *Pvu*II site in the 4-kilobase *crp*-containing *Bam*HI fragment of *E. coli* but two in the corresponding region of *Shigella flexneri* (data not shown). There was one site for *Ava*I on the same fragment in *E. coli* but none in *Shigella flexneri*. The *Hae*III and *Hin*fI digests were similar but not identical. Therefore we expected the *Shigella flexneri crp* gene to be located between the *Hin*dIII and *Eco*RV sites, as in *E. coli*. The sequence of the *Shigella flexneri crp* gene was performed by the strategy described in Fig. 1. For the plasmid pST280, we found a restriction map different from that of *E. coli* (Fig. 1). To localize the *crp* gene, we performed a partial *Sau*3A digestion of plasmid pST280 and cloned the resulting fragments into the *Bam*HI site of pBR322. We then analyzed the restriction map of the smallest recombinant plasmid (pST278) still conferring a CRP$^+$ Ap^r phenotype to a *crp* strain after transformation (Fig. 1). We then created deletions between the *Bst*EII, *Cla*I, *Nco*I or *Sal*I sites (see Materials and Methods). The deleted plasmids were no longer able to confer a CRP$^+$ phenotype on a *crp* strain. We therefore presumed *crp* to be around the *Bst*EII site located in the middle of the insert. This was confirmed by DNA sequencing in both directions from this site. Sequencing of the *crp* gene and its surrounding regions was then performed exclusively on plasmid pST280 (Fig. 1), as comparison of the restriction maps of pST280 and pST278 revealed in pST278 the presence (due to the subcloning strategy) of *Sau*3AI noncontiguous fragments on the chromosomal *crp* region.

Comparison of the *crp* gene sequences. The sequences of the *E. coli*, *Shigella flexneri*, and *S. typhimurium* genes are shown in Fig. 2. As expected from the restriction analysis, the nucleotide sequence of the *crp* gene of *Shigella flexneri* was almost identical to that of the *E. coli crp* gene. The deduced amino acid sequences were identical. Only 4 base pairs were different. One changed position 1 of codon 116 TTG → Leu for CTG. The 3 others changed position 3 of codons 149, 155, and 172.

We found a higher divergence between the nucleotide sequences of the *crp* genes of *S. typhimurium* and *E. coli*. The nucleotide sequences were 12.3% divergent. The nucleotide differences were not randomly distributed, with most of them occurring in the 3′ half of the gene. It is interesting that three of the four nucleotides differing in *E. coli* and *Shigella flexneri* were identical in *Shigella flexneri* and *S. typhimurium*. We found only one amino acid change, the replacement of Ala 118 by a serine. This residue was located in helix C. The three-dimensional structure of this region is well documented (29). Although several residues of helix C (residues 111 to 134) are involved in cAMP binding or subunit-subunit interaction, Ala 118 itself does not seem to

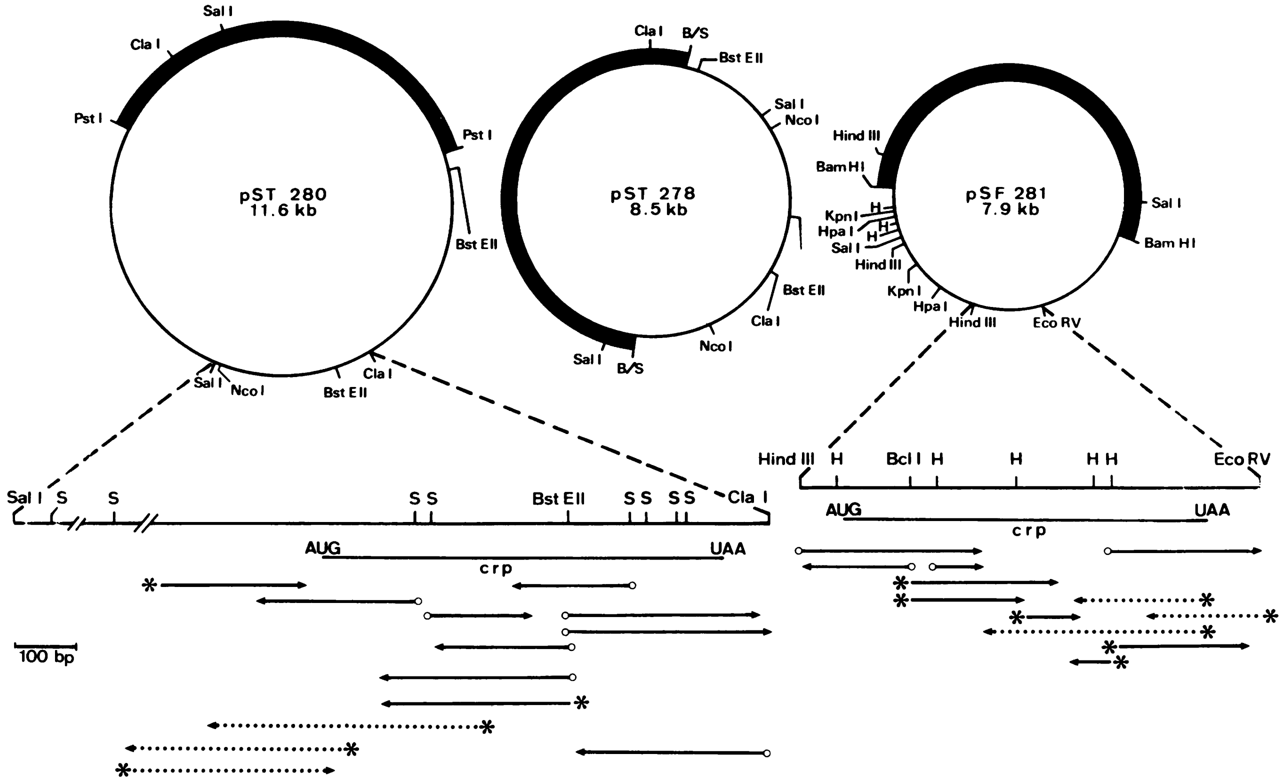

FIG. 1. Restriction maps of plasmids pST280, pST278, and pSF281 and sequencing strategy. The plasmids are represented to scale with their sizes indicated inside the circles (thick line, pBR322 DNA; thin line, bacterial insert). Restriction sites for enzymes having recognition sites of 4 or 5 base pairs: S, *Sau*3AI; H, *Hin*fI; Ha, *Hae*III. For pST280, *Sau*3AI sites are indicated only in the enlarged region, and for pSF281, only the *Hin*fI sites in the bacterial insert are shown. AUG——UAA, Coding region. →, sequences obtained by the Maxam and Gilbert technique; ....▸, sequences determined by the Sanger technique (see details in Materials and Methods). O and *, different strands.

play any particular role. In addition, it is interesting that this region of CAP is homologous to the regulatory subunit of the cAMP-dependent protein kinase from bovine cardiac muscle, which has a serine at this position (52). From these observations and our observation of identical phenotypes with both *crp* genes, one may assume that replacement of Ala 118 by a serine should not drastically affect the structure of CAP.

Comparison of the *crp* nucleotide sequences shows that *E. coli* and *Shigella flexneri* are more closely related to one another than either is to *S. typhimurium*. These results are in agreement with previous results obtained by classical taxonomical techniques (26, 50a), DNA-DNA hybridization of the bacterial chromosomes (7), or by comparison of the *trp* gene sequences (11, 35, 36, 53). We even found a lower degree of divergence between *Shigella flexneri* and *E. coli* than that observed between different *E. coli* strains. Milkman and Crawford (31) have reported considerably greater divergence in *trpB* gene sequences from different *E. coli* strains (up to about 10%) than we report here for the *crp* gene sequences for *E. coli* K-12 and *Shigella flexneri*.

Analysis of the diversity between *E. coli* and *S. typhimurium ompA*, *trp*, *metJ*, and *araC* genes has already been performed (3, 6, 8, 11, 18, 35, 36, 43, 48, 50, 51) (Table 1). Diversities of 9 to 25% for nucleotides and 3.5 to 15% for amino acids were found, with the nucleotide changes occurring principally in position 3 of a codon. *S. typhimurium* and *E. coli crp* genes show 12.3% diversity in their nucleotide sequences but only a 0.5% diversity in their amino acid sequences. Therefore CAP appears to be more highly conserved than are the proteins encoded by the *trp* operon, *ompA*, *metJ*, or *araC* genes in these two species.

Nature of nucleotide differences and codon usage in *S. typhimurium* and *E. coli crp* genes. The base pair differences between the *S. typhimurium* and *E. coli crp* genes are summarized in Table 2. All possible single base pair substitutions are represented; transitions were more common than were transversions. Although on a purely combinational basis transversions might have been expected to occur more frequently than transitions, analysis of the *trp* and *ompA* genes has already indicated that transitions were more abundant and accounted for 70% of the nucleotide changes. This was probably due to the stability of GT base pairs as intermediates during the GC → AT transitions and to the fact that the majority of transitions at position 3 of a codon did not lead to an amino acid change and are therefore favored over transitions affecting other bases in the codon. We have for the latter reason indicated the position of the changes in the codons in Table 2. Changes in position 3 are indeed favored over those in positions 1 and 2. We have included in Table 2 the values found for the nucleotide changes observed for *metJ* and *araC* genes. The latter seems to be an exception, with a high percentage of transversions accounting for at least part of the great number of amino acid differences.

We compared the pattern of codon usage in *crp* genes of *E. coli* and *S. typhimurium* (Table 3). Eleven codons (TTA, CTA, ATA, CCT, CCC, CGA, CGG, AGT, AGA, AGG, GGG) were never used in the *E. coli crp* gene. All these codons are rarely used in *E. coli* nonregulatory genes, in which the observed pattern of codon usage correlates

```
                     1                                                                           20
E. coli         ATG GTG CTT GGC AAA CCG CAA ACA GAC CCG ACT CTC GAA TGG TTC TTG TCT CAT TGC CAC ATT
S. flexneri
S. typhimurium                                                T

                    21                                                                           40
E. coli             CAT AAG TAC CCA TCC AAG AGC ACG CTT ATT CAC CAG GGT GAA AAA GCG GAA ACG CTG TAC
S. flexneri
S. typhimurium                    G   A               G                           A

                    41                                                                           60
E. coli             TAC ATC GTT AAA GGC TCT GTG GCA GTG CTG ATC AAA GAC GAA GAG GGT AAA GAA ATG ATC
S. flexneri
S. typhimurium                            C                           T       A   G

                    61                                                                           80
E. coli             CTC TCC TAT CTG AAT CAG GGT GAT TTT ATT GGC GAA CTG GGC CTG TTT GAA GAG GGC CAG
S. flexneri
S. typhimurium            T   T                                   T                           A

                    81                                                                          100
E. coli             GAA CGT AGC GCA TGG GTA CGT GCG AAA ACC GCC TGT GAA GTG GCT GAA ATT TCG TAC AAA
S. flexneri
S. typhimurium                C       C                           A       G   C               C

                    101                                                                         120
E. coli             AAA TTT CGC CAA TTG ATT CAG GTA AAC CCG GAC ATT CTG ATG CGT TTG TCT GCA CAG ATG
S. flexneri                                                                     C
S. typhimurium                        A   C       C           T               C C C     T C

                    121                                                                         140
E. coli             GCG CGT CGT CTG CAA GTC ACT TCA GAG AAA GTG GGC AAC CTG GCG TTC CTC GAC GTG ACG
S. flexneri
S. typhimurium            T       C T A           C   T   A       A   T       C   C       T       C   C

                    141                                                                         160
E. coli             GGC CGC ATT GCA CAG ACT CTG CTG AAT CTG GCA AAA CAA CCA GAC GCT ATG ACT CAC CCG
S. flexneri                                               C                       T
S. typhimurium            G   T   C   T       G                   G       G   C   T   C       G

                    161                                                                         180
E. coli             GAC GGT ATG CAA ATC AAA ATT ACC CGT CAG GAA ATT GGT CAG ATT GTC GGC TGT TCT CGT
S. flexneri                                                           C
S. typhimurium            T   G       G           C   T               T   C       C           C   C   C

                    181                                                                         200
E. coli             GAA ACC GTG GGA CGC ATT CTG AAG ATG CTG GAA GAT CAG AAC CTG ATC TCC GCA CAC GGT
S. flexneri
S. typhimurium                  T   T   T     T     A                   A                   G   T   C

                    201                             209
E. coli             AAA ACC ATC GTC GTT TAC GGC ACT CGT TAA TCCCGTCGGAGTGGCGCGTTACCTGGTAGCGCGCCATTT
S. flexneri
S. typhimurium            G               C       T   C          T     A  A         └┬┘ └┘  T │
                                                                                         ┌┴┐
                                                                                         AT

E. coli             TGTTT
S. flexneri
S. typhimurium
```

FIG. 2. Comparison of the coding regions of *E. coli*, *Shigella flexneri*, and *S. typhimurium crp* genes and their downstream region. The complete nucleotide sequence of the *E. coli crp* gene is shown. Only nucleotides which in *Shigella flexneri* or *S. typhimurium* are different from those in *E. coli* are indicated. └┬┘ and ┌┴┐, Deleted and inserted nucleotides.

roughly with the availability of isoaccepting tRNAs (23). However, it has been noticed that in infrequently expressed regulatory genes such as *lacI* (17), *araC* (48), *trpR* (45) or *dnaG* (46), "rare" codons are unusually highly used (24), and it has been proposed that codon usage might be a genome strategy for modulating gene expression (20). *crp*, which is more highly expressed (10) than is the *lacI*, *araC*, *trpR*, or *dnaG* gene, indeed did not use rare codons and showed a preference for highly expressed tRNAs. However, it is interesting that for phenylalanine, isoleucine, alanine, histidine, asparagine, and glycine, the *crp* gene did not use the most comonly used codon, a property which would preclude the placement of *crp* in the same category as very highly expressed genes like ribosomal proteins.

For the *S. typhimurium crp* gene, the general pattern of codon usage seemed to be the same as for *E. coli*: 8 of the 11 codons never used in *E. coli* were not used in the *S. typhimurium* gene either. The others (TTA, CCC, and GGG) were rarely used. In addition, CCA, GGA, and TCG, which were not used in the *S. typhimurium crp* gene, were rarely used in the *E. coli crp*. Differences in codon usage in both organisms were present, however, for isoleucine, valine, serine, proline, threonine, alanine, histidine, asparagine, and aspartic acid. These differences bore on a strong preference for a codon (for example, CCG for proline), a lack of preference (like equal use of CAT or CAC for histidine or of any of the four codons for alanine), or a completely different codon usage (as for aspartic acid or cysteine). In fact, it has already been observed for the *trp* structural genes that codon biases vary between *E. coli* and *S. typhimurium*. We also analyzed and detected differences in the codon usage in the *araC* and *metJ* genes of *E. coli* and *S. typhimurium*. But it seems that these differences bore essentially on a different use of codons, corresponding to abundant tRNAs, with the rare codons being roughly the same in both organisms (except for CCC, which is rare in *E. coli* but more abundant in *S. typhimurium*).

Comparison of the *crp* regulatory regions from *E. coli* and *S. typhimurium*. We determined the sequence of the 280 base pairs preceding the *crp* gene in *S. typhimurium* (Fig. 3). The initiation site for transcription has been determined for the *E. coli crp* gene (1). In addition, in vitro (1) and in vivo (10) analyses have shown that in *E. coli crp* is autoregulated. By footprinting experiments, Aiba (1) has shown the existence of two CAP binding sites: a specific high-affinity CAP binding site (CAP site I) located between the promoter and the coding region and another CAP binding site located upstream from the transcription start (CAP site II). Although the transcription start point of *crp* in *S. typhimurium* is not known, examination of the DNA sequence upstream from the structural gene revealed significant conservation in the regions presumed to be involved in the initiation of transcription and in its regulation. The sequence preceding *crp* in *E. coli* and in *S. typhimurium* shared a maximal homology after the introduction of two inserts of 1 base pair and one insert of 2 base pairs in the *S. typhimurium* sequence. The promoter region determined in *E. coli* was well conserved in *S. typhimurium*. The −10 and −35 regions of *E. coli* were absolutely identical in *S. typhimurium*. We observed two base pair changes in the region located between the transcription start site (+1) and the Pribnow box (−10 region). The sequence between the −10 and −35 regions is more A-T rich in *E. coli*, with two A's being changed to C or G, respectively, in *S. typhimurium*. A sequence identical to CAP site II was found at the homologous position. The CAP site I regions of the *S. typhimurium* and *E. coli crp* genes differed in two positions. One change was a C → T replacement at position 9, a position which was not conserved among the different CAP sites; the other was a T → C replacement at position 15, which was well conserved in the different CAP sites. CAP acts as a repressor for *crp* expression. The consensus sequence for the CAP sites (from 1 to 20, respectively) is A A-T G T G A-- T---T C A-A T. It has been determined (15) mostly from sites where CAP acts as an activator. These sites are oriented toward or opposite the direction of transcription. For CAP site I, it is opposite the direction of transcription. The presence of a change at a rather conserved position might indicate that the binding at sites where CAP acts as a repressor could be different from those where this protein acts as an activator. We also found that the region of translation initiation was well conserved: the 25 nucleotides before the ATG start codon and the 35 nucleotides following it were identical. Such a conserved nucleotide sequence (59 base pairs) is unusual. The complete homology between two sequences of 66 base pairs has been observed for *trpB* (11). In that case, the authors have proposed that it could be due to the occurrence of a genetic exchange between the two species. Such an event could have occurred for several parts of the chromosome, with the 59-base-pair conserved region of *crp* being one of them.

TABLE 1. Diversity between *E. coli* and *S. typhimurium* genes

Gene	% Diversity	
	Amino acid	Nucleotide
crp	0.5	12.3
metJ	2	5
araC[a]	8	18
trpE	12.5	20
trpG	4.1	18
trpB	3.5	16
trpA	15	25
ompA[a]	6.1	9

[a] Comparison was made on the common part. *araC* of *E. coli* has 292 codons, and that of *S. typhimurium* has 281; *ompA* of *E. coli* has 346 codons, and that of *S. typhimurium* has 350.

TABLE 2. Nature of nucleotide changes in *E. coli*→*S. typhimurium*

Change[a]	No. of base pair differences for gene[b]:		
	crp	*metJ*	*araC*
Transitions			
A → G	7	2	22
G → A	9	3	16
C → T	17	6	24
T → C	19	4	18
Transversions			
A → C	4	—[c]	10
A → T	3	1	2
G → C	7	—	10
G → T	3	—	11
C → A	2	—	8
C → G	1	—	9
T → A	1	1	9
T → G	5	1	13
Positions			
3	73	15	131
2	—	—	6
1	4	3	15

[a] Transitions accounted for 67.5, 83, and 53% of the changes for *crp*, *metJ*, and *araC*, respectively, whereas transversions represented 32.5, 17, and 47%, respectively.

[b] *araC* of *E. coli* has 292 codons; *araC* of *S. typhimurium* has 281. Comparison was made on the common part. For both species, *crp* has 210 codons and *metJ* has 105.

[c] —, No occurrence.

The highest region of diversity was in a transcribed but untranslated region downstream from CAP site I. In the region preceding the *crp* gene, interesting structural features were observed (see Fig. 2). The presence of two inverted

TABLE 3. Codon usage in *E. coli* and *S. typhimurium crp* genes compared with that in 25 *E. coli* genes

Amino acid and codon	Frequency (%) of usage			Amino acid and codon	Frequency (%) of usage		
	25 *E. coli* genes (% only)[a]	*E. coli crp*	*S. typhimurium crp*		25 *E. coli* genes (% only)	*E. coli crp*	*S. typhimurium crp*
Phe TTT	44	3 (60.0)	3 (60.0)	Tyr TAT	41	1 (16.7)	1 (16.7)
Phe TTC	56	2 (40.0)	2 (40.0)	Tyr TAC	59	5 (83.3)	5 (83.0)
Leu TTA	6.1	0 (0.0)	2 (9.1)	End TAA		1 (100)	1 (100)
Leu TTG	8	3 (13.6)	2 (9.1)	End TAG		0 (0.0)	0 (0.0)
Leu CTT	9	2 (9.1)	4 (18.2)	His CAT	39	2 (33.3)	3 (50.0)
Leu CTC	7	3 (13.6)	2 (9.1)	His CAC	61	4 (66.7)	3 (50.0)
Leu CTA	2	0 (0.0)	0 (0.0)	Gln CAA	27	5 (35.7)	4 (28.6)
Leu CTG	69	14 (63.6)	12 (54.5)	Gln CAG	73	9 (64.3)	10 (71.4)
Ile ATT	37	11 (64.7)	6 (35.3)	Asn AAT	24	2 (40.0)	2 (40.0)
Ile ATC	62	6 (35.3)	11 (64.7)	Asn AAC	76	3 (60.0)	3 (60.0)
Ile ATA	1	0 (0.0)	0 (0.0)	Lys AAA	77	12 (80.0)	12 (80.0)
Met ATG		7 (100.0)	7 (100.0)	Lys AAG	23	3 (20.0)	3 (20.0)
Val GTT	38	2 (14.3)	2 (14.3)	Asp GAT	51	2 (25.0)	6 (75.0)
Val GTC	13	3 (21.4)	7 (50.0)	Asp GAC	49	6 (75.0)	2 (25.0)
Val GTA	23	2 (14.3)	2 (14.3)	Glu GAA	73	13 (81.2)	15 (93.7)
Val GTG	27	7 (50.0)	3 (21.4)	Glu GAG	27	3 (18.7)	1 (6.3)
Ser TCT	27	4 (36.4)	4 (33.3)	Cys TGT	42	2 (66.7)	1 (33.3)
Ser TCC	26	3 (27.3)	5 (41.7)	Cys TGC	58	1 (33.3)	2 (66.7)
Ser TCA	8	1 (9.1)	1 (8.3)	End TGA		0 (0.0)	0 (0.0)
Ser TCG	11	1 (9.1)	0 (0.0)	TRP TGG		2 (100.0)	2 (100.0)
Pro CCT	9	0 (0.0)	0 (0.0)	Arg CGT	58	8 (72.7)	6 (54.5)
Pro CCC	6	0 (0.0)	1 (16.7)	Arg CGC	35	3 (27.3)	5 (45.5)
Pro CCA	20	2 (33.3)	0 (0.0)	Arg CGA	2	0 (0.0)	0 (0.0)
Pro CCG	65	4 (66.7)	5 (83.3)	Arg CGG	3	0 (0.0)	0 (0.0)
Thr ACT	24	5 (38.5)	2 (15.4)	Ser AGT	6	0 (0.0)	0 (0.0)
Thr ACC	51	4 (30.8)	6 (46.2)	Ser AGC	22	2 (18.2)	2 (16.7)
Thr ACA	6	1 (7.7)	1 (7.7)	Arg AGA	1	0 (0.0)	0 (0.0)
Thr ACG	20	3 (23.1)	4 (30.8)	Arg AGG	0.25	0 (0.0)	0 (0.0)
Ala GCT	28	2 (15.4)	3 (25.3)	Gly GGT	48	6 (37.5)	6 (37.5)
Ala GCC	19	1 (7.7)	3 (25.0)	Gly GGC	41	9 (56.2)	7 (43.7)
Ala GCA	23	6 (46.2)	3 (25.0)	Gly GGA	5	1 (6.3)	0 (0.0)
Ala GCG	30	4 (30.8)	3 (25.0)	Gly GGG	7	0 (0.0)	3 (18.7)

[a] Compiled by Konigsberg and Nigel-Godsen (24).

repeats was noted by Aiba et al. (2). Only one of them was completely conserved in *S. typhimurium*. We have noticed (unpublished data) an open reading frame (213 nucleotides) upstream from *crp* in *E. coli*. It was shorter in *S. typhimurium*. These observations suggest that neither the second inverted repeat nor the open reading frame seen in the *E. coli* sequence played a crucial role for *crp* expression, since we observed an identical phenotype for *crp* strains harboring plasmids containing the *crp* region from either organism. In addition to the differences in the CAP site, however, these differences suggest that *E. coli* and *S. typhimurium* might have used slightly different regulatory mechanisms not detected by the observation of a resulting phenotype on indicator plates, with genes present on a multicopy plasmid.

Other regulatory regions have been determined and compared for both *E. coli* and *S. typhimurium*. *araC* (22, 25, 33, 37), *ompA* (18, 34), and *trp* (for a review, see reference 12) regulatory regions from *S. typhimurium* and *E. coli*, respectively, exhibited very similar sequences for the regions shown to interact with the RNA polymerase (−10 or −35 regions) or with their respective regulatory proteins *araC*, CAP, or *trpR*; the sequences of the regions transcribed in a structured mRNA, such as the stems of the *trp* attenuator region, were conserved. Most of the changes occurred outside these different regions. We compared the recently described *metJ* regulatory regions from *E. coli* and *S. typhimurium* (43, 46a, 50) and also found that the changes were located outside the regions of DNA-protein interaction. Interestingly, as mentioned above for the *crp* gene, the region of highest diversity was immediately upstream from *metJ* and *araC* genes in the transcribed but untranslated region.

Comparison of sequences beyond *crp*. The nucleotide sequences of approximately 50 base pairs beyond the end of the *crp* genes of *E. coli*, *Shigella flexneri*, and *S. typhimurium* were determined (Fig. 2). Each sequence reveals the presence of structural features believed to be important in transcription termination (40): a G+C rich region followed by a U-rich sequence and a region of dyad symmetry that may form a hairpin structure in the mRNA. The *E. coli* and *Shigella flexneri* sequences were completely identical. The primary structures of *E. coli* and *S. typhimurium* downstream from the *crp* gene diverged somewhat, but the putative secondary structures of the transcript were conserved. However, the stems of the hairpins varied

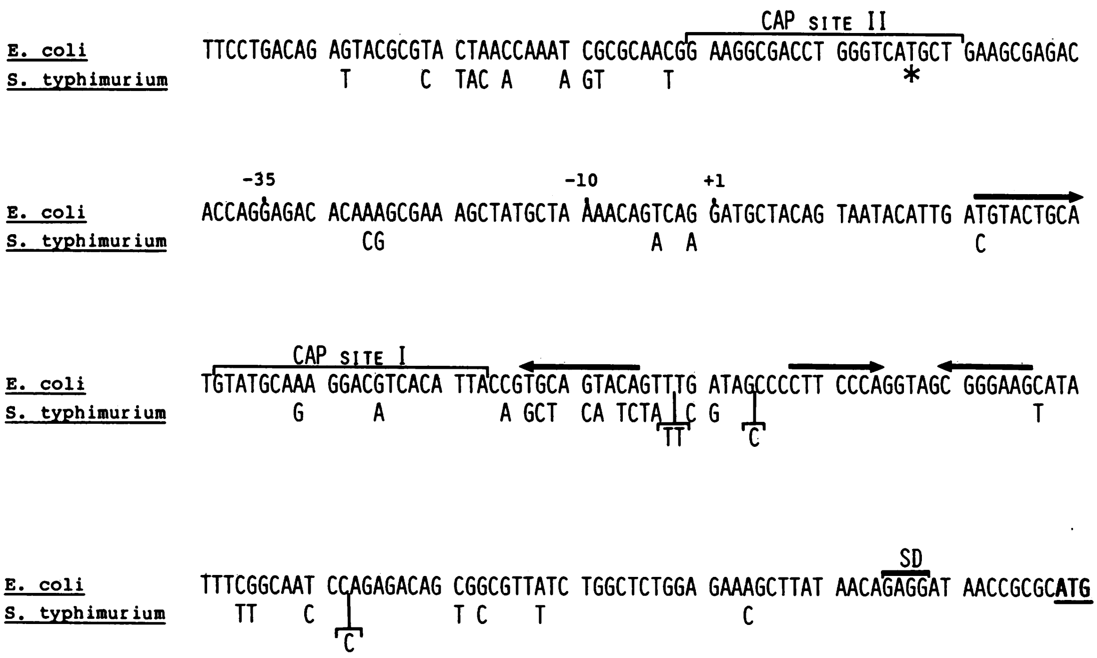
CAP SITE II
E. coli TTCCTGACAG AGTACGCGTA CTAACCAAAT CGCGCAACGG AAGGCGACCT GGGTCATGCT GAAGCGAGAC
S. typhimurium T C TAC A A GT T *
-35 -10 +1
E. coli ACCAGGAGAC ACAAAGCGAA AGCTATGCTA AAACAGTCAG GATGCTACAG TAATACATTG ATGTACTGCA
S. typhimurium CG A A C
CAP SITE I
E. coli TGTATGCAAA GGACGTCACA TTACCGTGCA GTACAGTTTG ATAGCCCCTT CCCAGGTAGC GGGAAGCATA
S. typhimurium G A A GCT CA TCTA C G T
TT C
SD
E. coli TTTCGGCAAT CCAGAGACAG CGGCGTTATC TGGCTCTGGA GAAAGCTTAT AACAGAGGAT AACCGCGCATG
S. typhimurium TT C T C T C
C

FIG. 3. Comparison of the regulatory regions of the *E. coli* and *S. typhimurium crp* genes. The nucleotide sequence of the regulatory region of the *E. coli crp* gene is shown. Only those nucleotides which are different in *E. coli* and *S. typhimurium* are indicated under the *E. coli* sequence. Additional nucleotides in *S. typhimurium* are indicated under ┌┴┐. The CAP (sites I and II) and RNA polymerase (−35 and −10) binding sites identified in *E. coli* (1) have been indicated. +1, Transcriptional initiation site determined for *E. coli* (1); ATG, translational start codon of *crp*; →, palindromic sequence detected in *E. coli* (see text and reference 12); □, translational start codon of a putative open reading frame in *E. coli* (see text).

in length and stability. *E. coli* seemed to have a longer and stabler hairpin than *S. typhimurium* had. For *S. typhimurium*, the stablest structure was obtained when part of the stretch of distal U's was paired with A's preceding the GC-rich region. In fact, a number of transcription terminators have been described recently that have a run of adenine residues preceding the GC-rich region, providing a symmetric counterpart to the U-encoding region (for a review, see reference 38), allowing the terminator to function in both directions. In the case of *S. typhimurium*, extending the stem and loop structure by including the U's did stabilize the secondary structure, which was not true for *E. coli*.

ACKNOWLEDGMENTS

We thank Wolf Epstein and Laszlo Csonka for their generous gift of strains. We thank Isabelle Sauvaget for help in using the computer facilities of the Unité Calcul de l'Institut Pasteur. We thank Georges Cohen for encouragement and constant interest in this work. We are endebted to Tony Pugsley for critical reading of the manuscript and very helpful discussion. We thank Lucile Girardot for secretarial assistance and Bénédicte Raymond for skillful drawing.

This work was supported by the Centre National de la Recherche Scientifique (ATP Microbiologie 84), contract 831005 from the Institut National de la Santé et de la Recherche Médicale, contract 83 V 0624 from the Ministère de l'Industrie et de la Recherche, and Public Health Service grant GM 29.067 from the National Institutes of Health. M.J.C. is the recipient of research career development award A 100468 from the National Institutes of Health.

LITERATURE CITED

1. **Aiba, H.** 1983. Autoregulation of the *Escherichia coli crp* gene: CRP is a transcriptional repressor for its own gene. Cell **32:**141–149.
2. **Aiba, H., S. Fujimoto, and N. Ozaki.** 1982. Molecular cloning and nucleotide sequencing of the gene for *E. coli* cAMP receptor protein. Nucleic Acids Res. **10:**1345–1362.
3. **Beck, E., and E. Bremer.** 1980. Nucleotide sequence of the gene *ompA* coding the outer membrane protein II* of *Escherichia coli* K-12. Nucleic Acids Res. **8:**3011–3027.
4. **Biggin, M. D., T. J. Gibson, and G. F. Hong.** 1983. Buffer gradient gels and ^{135}S label as an aid to rapid DNA sequence determination. Proc. Natl. Acad. Sci. USA **80:**3963–3965.
5. **Birnboim, H. C., and J. Doly.** 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res. **7:**1513–1523.
6. **Braun, G., and S. T. Cole.** 1982. The nucleotide sequence coding for major outer membrane protein OmpA of *Shigella dysenteriae*. Nucleic Acids Res. **10:**2367–2378.
7. **Brenner, D. J., G. R. Fanning, K. E. Johnson, R. V. Citarella, and S. Falkow.** 1969. Polynucleotide sequence relationships among members of *Enterobacteriaceae*. J. Bacteriol. **98:**637–650.
8. **Clarke, P., C. Lin, and G. Wilcox.** 1982. The nucleotide sequence of the *araC* regulatory gene in *Salmonella typhimurium* LT2. Gene **18:**157–163.
9. **Cossart, P., and B. Gicquel-Sanzey.** 1982. Cloning and sequence of the *crp* gene of *E. coli* K-12. Nucleic Acids Res. **10:**1363–1378.
10. **Cossart, P., and B. Gicquel-Sanzey.** 1985. Regulation of expression of the *crp* gene of *Escherichia coli* K-12: in vivo study. J. Bacteriol. **161:**454–457.
11. **Crawford, I. P., B. P. Nichols, and C. Yanofsky.** 1980. Nucleotide sequence of the *trpB* gene in *Escherichia coli* and *Salmonella typhimurium*. J. Mol. Biol. **142:**489–502.
12. **Crawford, I. P., and G. V. Stauffer.** 1980. Regulation of tryptophan biosynthesis. Annu. Rev. Biochem. **49:**163–195.
13. **Csonka, L. N., M. M. Howe, J. L. Ingraham, L. S. Pierson III, and C. L. Turnbough, Jr.** 1981. Infection of *Salmonella typhimurium* with coliphage Mu *d*l (Ap^r *lac*): construction of *pyr*::*lac* gene fusions. J. Bacteriol. **145:**299–305.

14. **de Crombrugghe, B., S. Busby, and H. Buc.** 1984. Activation of transcription by the cyclic AMP receptor protein, p. 129–167. *In* R. F. Goldberger and K. R. Yamamoto (ed.), Biological regulation and development, vol. 3B. Plenum Publishing Corp., New York.
15. **Ebright, R. H., P. Cossart, B. Gicquel-Sanzey, and J. Beckwith.** 1984. Mutations that alter the DNA sequence of the catabolite gene activator proteins of *E. coli.* Nature (London) **311**:223–235.
16. **Ebright, R. H., P. Cossart, B. Gicquel-Sanzey, and J. Beckwith.** 1984. Molecular basis of DNA sequence recognition by the catabolite gene activator proteins: detailed inferences from three mutations that alter DNA sequence specificity. Proc. Natl. Acad. Sci. USA **81**:7274–7278.
17. **Farabaugh, P. J.** 1978. Sequence of the *lacI* gene. Nature (London) **274**:765–769.
18. **Freudl, R., and S. T. Cole.** 1983. Cloning and molecular characterization of the *ompA* gene from *Salmonella typhimurium.* Eur. J. Biochem. **134**:497–502.
19. **Garges, S., and S. Adhya.** 1985. Sites of allosteric shift in the structure of the cyclic AMP receptor protein. Cell **41**:745–751.
20. **Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier.** 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9**(Suppl.):r43–r74.
21. **Groisman, E., B. A. Castilho, and M. J. Casadaban.** 1984. *In vivo* DNA cloning and adjacent gene fusion with a mini-Mu *c lac* bacteriophage containing a plasmid replicon. Proc. Natl. Acad. Sci. USA **81**:1480–1483.
22. **Horwitz, A. H., L. Heffernan, C. Morandi, J. Lee, J. Timko, and C. Wilcox.** 1981. DNA sequence of the *araBAD*-araC controlling region in *Salmonella typhimurium* LT2. Gene **14**:309–319.
23. **Ikemura, T.** 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. **146**:1–21.
24. **Konigsberg, W., and G. Nigel-Godson.** 1983. Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia. coli.* Proc. Natl. Acad. Sci. USA **80**:687–691.
25. **Lee, N. L., W. O. Gielow, and R. G. Wallace.** 1981. Mechanism of araC autoregulation and the domains of two overlapping promoters, Pc and P_{BAD} in the L-arabinose regulatory region of *Escherichia coli.* Proc. Natl. Acad. Sci. USA **78**:752–756.
26. **Le Minor, L.** 1982. Enterobacteries, p. 240–315. *In* L. Le Minor and M. Véron (ed.), Bactériologie médicale. Flammarion, Medicine Sciences, Paris.
27. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1983. Molecular cloning: a laboratory manual, p. 545. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
28. **Maxam, A. M., and W. Gilbert.** 1977. A new method for sequencing DNA. Proc. Natl. Acad. Sci. USA **74**:560–564.
29. **McKay, D., I. Weber, and T. Steitz.** 1982. Structure of catabolite gene activator protein at 2.9 Å resolution. J. Biol. Chem. **257**: 9518–9524.
30. **Messing, J., and J. Vieira.** 1982. A new pair of M13 vectors for selecting either DNA strand of double digest restriction fragments. Gene **19**:269–276.
31. **Milkman, R., and I. P. Crawford.** 1983. Clustered third-base substitutions among wild strains of *Escherichia coli.* Science **221**:378–380.
32. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
33. **Miyada, C. C., A. H. Horwitz, L. G. Cass, J. Timko, and G. Wilcox.** 1980. DNA sequence of the *araC* regulatory gene from *Escherichia coli* B/r. Nucleic Acids Res. **8**:5267–5274.
34. **Movva, R. N., K. Nakamura, and M. Inouye.** 1980. Regulatory region of the gene for the OmpA protein, a major outer membrane protein of *Escherichia coli.* Proc. Natl. Acad. Sci. USA **77**:3845–3849.
35. **Nichols, B., G. Miozzari, M. Van Cleemput, G. N. Bennett, and C. Yanofsky.** 1980. Nucleotide sequences of the trpG regions of *Escherichia coli, Shigella dysenteriae, Salmonella typhimurium* and *Serratia marcescens.* J. Mol. Biol. **142**:503–517.
36. **Nichols, B., and C. Yanofsky.** 1979. Nucleotide sequences of trpA of *Salmonella typhimurium* and *Escherichia coli*: an evolutionary comparison. Proc. Natl. Acad. Sci. USA **76**:5244–5248.
37. **Ogden, S., D. Haggerty, C. Stoner, D. Kolodrubetz, and R. Schleif.** 1980. The *Escherichia coli* L-arabinose operon: binding sites of the regulatory proteins and a mechanism of positive and negative regulation. Proc. Natl. Acad. Sci. USA **77**:3346–3350.
38. **Postle, K., and R. F. Good.** 1985. A bidirectional rho independent-transcription terminator between the *E. coli tonB* gene and opposing gene. Cell **41**:577–585.
39. **Rigby, P. W. J., M. Dieckmann, C. Rhodes, and P. Berg.** 1977. Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J. Mol. Biol. **113**:237–251.
40. **Rosenberg, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13**:319–353.
41. **Rothman-Denes, L. B., J. E. Heese, and W. Epstein.** 1973. Role of cyclic adenosine 3′,5′-monophosphate in the in vivo expression of the galactose operon of *Escherichia coli.* J. Bacteriol. **114**:1040–1044.
42. **Sabourin, D., and J. Beckwith.** 1975. Deletion of the *Escherichia coli crp* gene. J. Bacteriol. **122**:338–340.
43. **Saint-Girons, I., N. Duchange, G. N. Cohen, and M. Zakin.** 1984. Structure and regulation of the *metJ* regulatory gene in *Escherichia coli.* J. Biol. Chem. **259**:14282–14285.
44. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
45. **Singleton, C. K., W. D. Roeder, G. Bogosian, R. L. Somerville, and H. L. Weith.** 1980. DNA sequence of the *E. coli trpR* gene and prediction of the amino acid sequence of Trp repressor. Nucleic Acids Res. **8**:1551–1560.
46. **Smiley, B. L., J. R. Lupski, P. S. Svec, R. McMaken, and G. N. Godson.** 1982. Sequences of the *Escherichia coli dnaG* primase gene and regulation of its expression. Proc. Natl. Acad. Sci. USA **79**:4550–4554.

46a. **Smith, A. A., R. C. Greene, T. W. Kirby, and B. R. Hindenach.** 1985. Isolation and characterization of the methionine regulatory gene, *metJ*, of *Escherichia coli* K-12. Proc. Natl. Acad. Sci. USA **82**:6104–6108.

47. **Southern, E.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98**:503–517.
48. **Stoner, C. M., and R. Schleif.** 1982. Is the amino acid but not the nucleotide sequence of the *Escherichia coli araC* gene conserved? J. Mol. Biol. **154**:649–652.
49. **Ullmann, A., and A. Danchin.** 1983. Role of cyclic AMP in bacteria. Adv. Cyclic Nucleotide Res. **15**:1–53.
50. **Urbanowski, M. L., and G. V. Stauffer.** 1985. Nucleotide sequence and biochemical characterization of the *metJ* gene from *Salmonella typhimurium* LT2. Nucleic Acids Res. **13**:673–681.

50a. **Véron, M.** 1982. Taxonomie bacterienne, p. 80–97. *In* L. Minor and M. Véron (ed.), Bactériologie médicale. Flammarion Medicine Sciences, Paris.

51. **Wallace, R. G., N. Lee, and A. Fowler.** 1980. The *araC* gene of *Escherichia coli*: transcriptional and translational start points and complete nucleotide sequence. Gene **12**:179–190.
52. **Weber, I. T., K. Takio, K. Titani, and T. A. Steitz.** 1982. The cAMP binding domains of the regulatory subunit of cAMP-dependent protein kinase and the catabolite gene activator protein are homologous. Proc. Natl. Acad. Sci. USA **79**: 7679–7683.
53. **Yanofsky, C., and M. Van Cleemput.** 1982. Nucleotide sequence of type of *Salmonella typhimurium* homology with the corresponding sequence of *Escherichia coli.* J. Mol. Biol. **154**:235–246.